Anonimizálási gyakorlat? – Egy magyar korpusz anonimizálásának tanulságai

Mátyus Kinga

MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
matyus.kinga@nytud.mta.hu

Kivonat: A Budapesti Szociolingvisztikai Interjú (BUSZI) második változata, vagyis a BUSZI-2 mintegy 100 órányi felvétele számos kutatásra ad lehetőséget. Az interjúk irányított beszélgetései sok személyes adatot tartalmaznak, melyek segítségével az adatközlők azonosíthatók. Azonban ahhoz, hogy a kutatók a korpuszhoz hozzáférjenek, az abban lévő érzékeny adatokat kezelni kell. A BUSZI-2 ötven interjúját az adatvédelmi törvény (1992. évi LXIII.) alapján, a korpusz jellegzetességeit is szem előtt tartva, illetve a nemzetközi gyakorlatra építve anonimizáltuk.

1 Bevezetés

Az MTA Nyelvtudományi Intézet Élőnyelvi Munkacsoportja 1985-ben kezdte meg egy olyan nagyszabású, beszélt nyelvet vizsgáló, abban az időben nemzetközi szinten is kimagasló kutatás megtervezését, melynek célja az volt, hogy az eddigi írott korpuszokon alapuló leírásokat jelentős nagyságú beszélt nyelvi korpusz elemzéséből nyert adatokkal kiegészíthessék, módosíthassák (részletesen lásd [5]).

A Budapesti Szociolingvisztikai Interjú (BUSZI) elnevezésű projektnek 2 korpusza készült el. A BUSZI-2 felvételeit 1987-ben rögzítették 50 adatközlővel, kvótaminta alapján: 5 különböző foglalkozási csoport 10-10 adatközlőjével készítettek 1,5–2,5 órás felvételeket. A BUSZI-3,-4 200 adatközlővel készült, rétegzett mintavétel alapján [8]. A korpuszok OTKA és AKP pályázati támogatásokkal[1] jöttek létre, a projektum vezetője Kontra Miklós volt. A BUSZI-2 ma már regisztrált kutatók számára hozzáférhető,[2] a BUSZI-3, -4 feldolgozása folyamatban van.

Mind a BUSZI-2, mind a BUSZI-3, -4 két nagy részből állt: kártyás feladatokból/ tesztfeladatokból és irányított beszélgetésekből.[3] Az irányított beszélgetések a Labov [6] által kidolgozott módszert követve az adatközlőknek nem megkomponált interjúnak, hanem inkább oldott beszélgetésnek tűntek, ennek is köszönhető, hogy rendkívül sok személyes adatot tartalmaznak. Ahhoz, hogy ehhez a korpuszhoz bármilyen kutató

[1] Részletesen lásd: http://buszi.nytud.hu/a-buszi-rol
[2] http://buszi.nytud.hu/kutatni-szeretnem-a-buszi-t
[3] Részletesen lásd: http://buszi.nytud.hu/a-buszi-rol/az-interjuk-felepitese

hozzáférjen, a benne található személyes adatokat kezelni kell. Jelen tanulmány a BUSZI-2 korpusz anonimizálását mutatja be.

1.1 Jogi szabályozás Magyarországon

A személyes adatok védelméről és a közérdekű adatok nyilvánosságáról Magyarországon az 1992. évi LXIII. törvény rendelkezik. Ennek értelmében a BUSZI-2 mind személyes (név, azonosító jel, fizikai, fiziológiai jellemzők stb.), mind pedig különleges adatokat (faji eredet, kisebbség, vallás, egészségi állapot stb.) tartalmaz. A kutathatósághoz szükséges, hogy az adatok és az érintett közti kapcsolat helyreállíthatóságát megszüntessük.

2 Anonimizálás más korpuszokban

Azok a beszélt nyelvi korpuszok, melyek azzal a céllal készültek, hogy szélesebb közönség számára elérhetőek legyenek, a következő gyakorlatokat követik: hozzájárulási nyilatkozatot kérnek az adatközlőktől, hogy a korpuszt közzé lehessen tenni. Az írott szövegek anonimitása könnyen biztosítható azzal, hogy törlik, vagy megváltoztatják például a neveket a korpuszban, azonban a hangfájlok anonimizálása már számos vélemény szerint nem lehetséges [4]. A Routledge Handbook of Corpus Linguistics szerzői hasonló alapelveket fogalmaznak meg: a tradicionális megközelítés szerint az adatközlők anonimitását hangsúlyosan szem előtt kell tartani. Ehhez a beszélők nevét és más részleteket meg kell változtatni, vagy teljesen törölni kell. Az anonimizálás bizonyos gyakran használt szavakra, kifejezésekre, sőt témákra is kiterjedhet, amelyek bármilyen módon felismerhetővé tehetik az adatközlőt. Az anonimitás még problematikusabb a hang- és videofelvételek esetén. Az egyéni hang, mint egy ujjlenyomat, azonosítja a beszélőt. A hang eltorzítása azonban akadályozhatja a korpusz fonetikai/fonológiai kutathatóságát – ezért nem ajánlott [1]. A személyes adatok kezelését szabályozó törvények minden országban különbözőek.

Lou Bernard a British National Corpus (BNC) anonimizálása kapcsán a következő lehetőségeket vázolta: 1) a nevet egyszerűen törölni kell, vagy XXXX-szel helyettesíteni, 2) egy kódot kell alkalmazni, amely minden Maggie betűsort egy kódra (XYZ12-re) cserél, vagy 3) egy szótárban a hasonló neveket hasonlóan fordítják – pl. a Maggie-nek Susan, a Jonesnak Brown lehetne a megfelelője [3].

A következőkben röviden bemutatjuk néhány korpusz gyakorlatát.

2.1 British National Corpus (BNC)

Azzal a céllal készült a korpusz, hogy a gyűjtött anyagot széles körben elérhetővé tegyék. Teljes anonimitást és bizalmas adatkezelést ígértek az adatközlőknek. (A beszélt részben, amely a korpusz 10%-át jelenti, 1) az adatközlők mikrofont viseltek, és felvették saját beszélgetéseiket, illetve 2) terepmunkások különböző műfajú beszédeket rögzítettek.)

A neveket és címeket törölték a korpuszból és a kapcsolódó dokumentumokból, helyette a <gap> címke áll magyarázattal, pl. <gap desc="name" reason= "anonymization"/>. Azt az ötletet, hogy a neveket kóddal vagy egy nyelvileg hasonló névvel helyettesítsék, praktikussági szempontok miatt elvetették.

2.2 Newcastle Electronic Corpus of Tyneside English (NECTE)

Amikor a BUSZI projektum az 1980-as évek végén elkezdődött, a ma is hatályos adatvédelmi törvény még nem élt. Hasonló volt a helyzet a Newcastle Corpus of Tyneside English (NECTE) esetében is. Beal [2] beszámol arról, milyen etikai és jogi feladatokat kellett megoldaniuk egy olyan korpusz közzététele során, amely még az 1998-as brit adatvédelmi törvény előttről származik. A NECTE egy hagyatékkorpusz, amely két részből áll: Tyneside Linguistic Survey (TLS) (1969-ből), és a Phonological Variation and Change Project (PVC) (1994-ből). A projekt célja az volt, hogy minél szélesebb körnek elérhetővé tegyék e két korpuszt. Mivel a TLS korpusz adatközlőinek anonimitást ígértek, a kutatók minden nevet töröltek a felvételekről és az átiratokról [2].

A TLS interjúk „érzékeny" témákat is érintenek (egészség, vallás, politika, szakszervezet), nem lenne elfogadható az a megoldás, hogy a korpuszt az interneten teszik hozzáférhetővé. Ezért azoknak a kutatóknak, akik a NECTÉ-vel szeretnének dolgozni, ki kell tölteniük egy nyomtatványt, s aláírva vissza kell küldeniük a központba.

2.3 BEA – magyar spontánbeszéd-adatbázis

A magyar Beszélt Nyelvi Adatbázis (BEA) esetében, az Amerikai Egyesült Államokbeli gyakorlathoz hasonlóan az adatközlőknek alá kell írniuk egy hivatalos hozzájárulási nyilatkozatot, és a kutatás során a személyes adatokat az interjútól külön kezelik, a kutatók a személyes adatokhoz nem juthatnak hozzá – kivéve természetesen a kutatáshoz szükséges adatokat, mint például a magasság és a kor.

3 A BUSZI-2 anonimizálása

A BUSZI-2 esetében az összes interjút kódolták, illetve lejegyezték, és az átiratokat kétszer ellenőrizték, hangfelvételek digitalizálták. Oravecz Csaba és Sass Bálint a szöveges lejegyzésből nyelvi adatbázist készítettek [7]. Az XML-fájl a lejegyzett beszélgetések elemzett változatát tartalmazza, amelyben megtalálhatóak a következő információk: egyértelműsített morfológiai elemzés, szótő; a regularizált szótő CV váza és a magánhangzók BNF (back, neutral, front) alakban; valamint az elhangzott szóalak fonetikai reprezentációja.

Az adatvédelmi törvény legfontosabb előírásai: megszüntetni az adatok és az érintettek közti kapcsolatot az érzékeny adatok törlésével az átiratokban, illetve kisípolásával a hangfelvételekben, emellett az egész hangfelvétel eltorzítása. Emellett szem előtt tartottuk azt is, hogy a BUSZI szociolingvisztikai interjúként csak akkor tud jól

funkcionálni, ha a beszélők egyes kutatási szempontból szükséges jellemzőit meghagyjuk – ilyen fontos jellemző például a születési hely, illetve az a hely, ahol az adatközlő a gyermekkorát töltötte. Az ilyen, szociolingvisztikai szempontból kiemelkedően fontos adatokat tehát meghagytuk, ám a beszélő anonimitását így is igyekeztünk biztosítani. Az a cél vezérelt bennünket, hogy minden tulajdonnév mechanikus törlése, illetve kisípolása helyett csak annyi, az adatközlőre utaló információt töröljünk, amelyek meghagyása veszélyeztette volna az adatközlő anonimitását.

A BUSZI-2 korpusz minden interjúja egyedi, ezért nem alkalmaztunk egységes szabályokat az anonimizálás során, hanem minden interjú adatait egyenként szűrtünk. A tesztfeladatok esetében csak a hangfelvételek torzítására volt szükség. Ehhez a SoX 14.4.0 programot használtuk.[4] A torzítás mértékét minden esetben az adatközlő hangjához igazítottuk. Az irányított beszélgetésekben minden esetben töröltük az adatközlő nevét és születési dátumát, és jellemzően töröltük a következő adatokat: munkahely címe, adatközlő jelenlegi és korábbi iskoláinak neve, az adatközlő lakhelye (utca). Nem töröltük viszont általában a következőket: születési hely, az adatközlő rokonainak adatai (kivétel a teljes név), az adatközlő lakhelye (kerület/városrész).

A doc formátumú átiratokban az érzékeny adatok helyén egy címke szerepel a törölt adat jellegével (pl. utcanév, személynév), az XML-formátumban MASKED címke szerepel a törölt elem helyén. A hangfájlokban a törölt elemek helyére azonos hosszúságú sinusjelet szúrtunk be, végül a teljes felvételt torzítottuk.

Egy példa anonimizált adatra a doc fájlban:

tm: december kilenc ■ öö kollégium a színhelye ■ az interjúnak. Budapesti interjú kvóta. ■ Be akarja mondani a nevét mar= vagy a marad a ka= marad inkább névtelenül?

ak: Jó mondom, mondhatom.

tm: Akko tessék.

ak: [#szemelynev]nak hívnak.

4 A BUSZI-2 kutathatóságáról

A BUSZI-2 ma már regisztrált kutatók számára hozzáférhető az interneten.[5] A tesztfeladatok szabadon elérhető eredményeinek feldolgozását egy keresőprogram segíti, melynek segítségével kilistázhatjuk, illetve meg is hallgathatjuk a részleteket (torzított formában).[6] Az irányított beszélgetések anonimizált átiratait doc, pdf és XML-formátumban tölthetik le a kutatók. Az ötven interjú átiratainak kutatását is egy keresőprogram segíti.[7] Az anonimizált és torzított hangfájlokat (a NECTÉ-hez és BNC-hez hasonlóan) nem tettük interneten elérhetővé, azok meghallgatására a Nyelvtudományi Intézetben előzetes időpont-egyeztetés után van lehetőség.

[4] http://sox.sourceforge.net/

[5] Részletesen lásd: http://buszi.nytud.hu/kutatni-szeretnem-a-buszi-t

[6] A tesztfeladatok keresőprogramját Blága Szabolcs készítette.

[7] Az irányított beszélgetések keresőprogramját Sass Bálint készítette.

Bibliográfia

1. Adolphs, S., Dawn, K.: Building a spoken corpus: What are the basics? In: O'Keeffe, A., McCarthy, M. (eds.): The routledge handbook of corpus lingusitics. Routledge, London (2010) 38–52
2. Beal, J. C.: Creating corpora from spoken legacy material. In: Renouf, A., Kehoe, A. (eds.): Corpus linguistics: Refinements and reassessments. Rodopi B. V., Amsterdam (2009) 33–47
3. Burnard, L.: A Note on Anonymization. Elérhető: http://www.natcorp.ox.ac.uk/archive/vault/pcw47.txt
4. Hunston, S.: Collection strategies and design decisions. In: Lüdeling, A., Merja, K. (eds.): Corpus linguistics: An international handbook. Volume 1. Mouton de Gruyter, Berlin (2008) 154–167
5. Kontra M.: Budapesti élőnyelvi kutatások. Magyar Tudomány, Vol. 5 (1990) 512–520
6. Labov, W.: Field methods of the project on linguistic change and variation. In: Baugh, J., Sherzer, J. (eds.): Language in use: Readings in sociolinguistics. Prentice-Hall, Englewood Cliffs, N. J. (1984) 28–53
7. Oravecz Cs., Sass B.: Szöveges lejegyzésből nyelvi adatbázis. Előadás. Elhangzott: BUSZI I. szimpózium, Budapest, 2008. december 9. Elérhető: http://www.nytud.hu/oszt/korpusz/resources/ocs_sb_buszidb.pdf
8. Váradi T.: A Budapesti Szociolingvisztikai Interjú. In: Kiefer, F. (szerk.): A magyar nyelv kézikönyve. Akadémiai Kiadó, Budapest (2003) 339–360